INTERIM TECHNICAL REPORT

PROJECT A–1621

STUDY OF USGS/NASA LAND USE CLASSIFICATION SYSTEM

By
G. William Spann and N.L. Faust

[redacted]

Prepared for

National Aeronautics and Space Administration
George C. Marshall Space Flight Center
Marshall Space Flight Center, Alabama

December 1974

ENGINEERING EXPERIMENT STATION
Georgia Institute of Technology
Atlanta, Georgia 30332

STUDY OF USGS/NASA LAND-USE CLASSIFICATION SYSTEM

INTERIM TECHNICAL REPORT

by

G. W. Spann and N. L. Faust

December, 1974

Prepared for

George C. Marshall Space Flight Center
Marshall Space Flight Center, Alabama 35812

Under Contract Number NAS8-30653

by

Engineering Experiment Station
Georgia Institute of Technology
Atlanta, Georgia 30332

ABSTRACT

It is known from several previous investigations that many categories of land-use can be mapped via computer processing of Earth Resources Technology Satellite Data. This report presents the results of one such experiment using the USGS/NASA land-use classification system.

Douglas County, Georgia, was chosen as the test site for this project. It was chosen primarily because of its recent rapid growth and future growth potential.

Results of the investigation indicate an overall land-use mapping accuracy of 67% with higher accuracies in rural areas and lower accuracies in urban areas. It is estimated, however, that 95% of the State of Georgia could be mapped by these techniques with an accuracy of 80% to 90%.

ACKNOWLEDGEMENTS

The authors would like to thank Mr. Bruce Rado of the Georgia Department of Natural Resources for his help in selecting the test site and evaluating the results of our efforts. We would also like to express appreciation for the assistance in gathering ground truth data provided by Ms. Kay Marsolan, Douglas County Planner, and Mr. Mike Swain, Acting County Engineer, also from Douglas County.

TABLE OF CONTENTS

I. INTRODUCTION

Background

From the results of several previous investigations by various groups it is obvious that land-use can be mapped via computer processing of Earth Resources Technology Satellite (ERTS) data [1,2,3,4]. However, many of the projects carried out to date have been special purpose in the sense that they were either very specifically directed toward one goal, or alternatively any land-use categories that fell out were mapped. In one project, for example, a land-use map of Milwaukee County was prepared which had five categories of water displayed. None of the above is meant to criticize the results of previous studies; however, it is intended to point out the lack of uniformity resulting from many previous land-use investigations using computer processing of ERTS data.

There is at the present time intense interest in and support for enactment of a national land-use bill. Should passage of this bill eventually take place, there is considerable merit in using a national land-use classification scheme for any mapping carried out under this proposed legislation. One such system has been proposed by James R. Anderson, et al., specifically for use with remote sensor data [5]. The categories of land-use proposed are given in Figure 1. As can be seen there are two levels of classification with Level II being a finer categorization of the Level I land-use classes.

As stated in the publication, Level I classifications were derived so that the source of information could be "satellite imagery, with very little supplemental information." The sources of information required for Level II were expected to be "high-altitude and satellite imagery combined with topographic maps." Several investigations have shown, however, that it is possible to map many categories in Level II directly from the ERTS data tapes (with appropriate ground truth information). Due to the varied nature of these investigations, it is difficult to identify all of the Level

Land-Use Classification System for Use With Remote Sensor Data

Level I	*Level II*
01. Urban and Built-up Land.	01. Residential.
	02. Commercial and services.
	03. Industrial.
	04. Extractive.
	05. Transportation, Communications, and Utilities.
	06. Institutional.
	07. Strip and Clustered Settlement.
	08. Mixed.
	09. Open and Other.
02. Agricultural Land.	01. Cropland and Pasture.
	02. Orchards, Groves, Bush Fruits, Vineyards, and Horticultural Areas.
	03. Feeding Operations.
	04. Other.
03. Rangeland.	01. Grass.
	02. Savannas (Palmetto Prairies).
	03. Chaparral.
	04. Desert Shrub.
04. Forest Land.	01. Deciduous.
	02. Evergreen (Coniferous and Other).
	03. Mixed.
05. Water.	01. Streams and Waterways.
	02. Lakes.
	03. Reservoirs.
	04. Bays and Estuaries.
	05. Other.
06. Nonforested Wetland.	01. Vegetated.
	02. Bare.
07. Barren Land.	01. Salt Flats.
	02. Beaches.
	03. Sand Other Than Beaches.
	04. Bare Exposed Rock.
	05. Other.
08. Tundra.	01. Tundra.
09. Permanent Snow and Icefields.	01. Permanent Snow and Icefields.

Figure 1. USGS/NASA Land-Use Classification System.

II categories which can or can not be mapped utilizing computer processing of ERTS data.

Present Program

In order to provide a consistent basis for discussing land-use mapping via ERTS, the present program was instituted. The general objective of this program is, thus, a determination of the extent to which the USGS/NASA land-use classification system is compatible with the computer processing techniques employed for land-use mapping from ERTS data. However, there are additional objectives to this program. The first is an assessment of the adequacy of this type of land-use mapping for meeting the needs of agencies responsible for land-use planning. A second objective is a cost-effectiveness study detailing the advantages/disadvantages of this methodology of land-use mapping over manual methods.

One of the current problems facing land-use planners is lack of a common vocabulary with the specialists who process remote sensing data. The USGS/NASA land-use classification system is an attempt to bridge this communication gap. However, there is still some confusion because automatic processing is capable of identifying more categories than those contained in Level I but less categories than are contained in Level II. At the conclusion of this study, it is anticipated that it will be possible to specify those categories of land-use which can be identified using ERTS data. This should provide a common ground on which land-use planners and processing specialists can begin working together to solve land-use problems.

Yet another parameter to be derived from this effort is a measure of the cost-effectiveness of automated land-use analyses. The data from the study will allow an estimation of the costs and benefits to be derived from the use of ERTS data for large scale land-use analysis efforts. These will be compared and contrasted with presently used manual methods of analysis, and with other estimates of costs given in the published literature.

The Georgia Department of Natural Resources (DNR) has agreed to participate in the study by providing inputs on the applicability of these results to operational planning agencies. In addition, DNR plans to supply cost data derived from other land-use mapping projects. Since Douglas County was chosen as the test site for this project, the Douglas County Planning Office has agreed to provide inputs necessary to the study. Other planning agencies will also be asked to provide advice and criticism pertinent to the results of this project. The reasons for choosing Douglas County are outlined in Section II.

Results to Date

The results contained in this report cover the first six months efforts on this project. While some of the results may be modified somewhat by later work, no major changes are anticipated. The results achieved to date were deemed sufficiently important that this report is being prepared in addition to the regularly scheduled reports specified for this project.

While the processing of ERTS data on the test site will continue throughout the project, preliminary conclusions can be drawn from the first six months work. All Level I categories are separable in the computer processed ERTS data. We have also been able to identify those categories in Level II which are separable in the ERTS data and those which overlap with other categories. A complete discussion of this topic is contained in Section III.

To check the accuracy of the computer generated land-use maps, NASA high altitude photographs and low altitude photographs were obtained, and field checks were carried out. This portion of the project is discussed in detail in Section IV. Section V contains some unexpected geological/soil association results from this project. It was found that vegetation cover provides an excellent indication of geology and soil types along the Brevard Fault zone in Douglas County.

Section VI contains a discussion of some philosophical issues raised by the results of this project. It also contains some proposed techniques for additional ERTS data processing. A summary of the results to date and conclusions to be drawn from these results are outlined in Section VII.

II. REASONS FOR CHOOSING DOUGLAS COUNTY TEST SITE

Douglas County is at an earlier stage in its development than many counties in the Metro Atlanta area (see Figure 2). However, several recent and pending events promise to accelerate rapidly the growth of this area. Of necessity this means that land-use patterns are changing rapidly and will continue to do so in the future. It is important, therefore, in this county that there be planning for the impacts on land-use which will occur. For these reasons, Georgia DNR selected Douglas County as an appropriate test site (see Figure 3).

The single major cause of the county's present rapid growth in residential and other areas is the recent completion of Interstate 20 into the county. This provides relatively easy access to the area from the center of Atlanta. As usually happens with the opening of a new transportation corridor, many families have chosen to locate along I-20 in Douglas County. Since I-20 presently ends within the county, many people who might otherwise live further from the center of Atlanta, probably locate in Douglas County. For whatever reasons, the recent completion of I-20 into the county seems to have accelerated the growth of the county (see Figure 4).

Pending events could have a much greater impact on Douglas County than simple outward growth from Atlanta. A site in the north portion of Douglas County is one of the proposed locations for a second Atlanta airport. If this should occur, many new industrial, commercial, and residential areas will open up within the county. One logical transportation corridor to the airport site would be a limited access highway originating at I-20 in Douglas County and terminating at the new airport. This would further increase pressures for development in Douglas County.

A west Georgia tollway has been proposed to link Chattanooga with Tallahassee. Should this road be built it would pass through or near the western portion of Douglas County. This major North-South transportation route would certainly impact the development of the west Georgia area, including the Douglas County area.

Location: 25 miles west of Atlanta.

Highways Serving: U.S. 78; Georgia 92, 166 and 5; Interstate 20 (east-west).

Population:

	1960 Census	1970 Census	1973 Estimate[1/]
Douglasville	4,462	5,472	6,500
Douglas County	16,741	28,659	44,509

Labor Force Estimate - Douglas County: (Georgia Department of Labor, June 1973)

Civilian labor force	5,070
Employed	4,450
Employed in manufacturing	590
Unemployment	620

Largest Manufacturers:

Company	Product	Employees
DeSoto Falls Spinners, Inc.	Synthetic yarns	134
Timms Mills, Inc.	Polyester yarns	125
Southern Empire Egg Farm	Egg Processing	45

Transportation:

Motor Freight - Barnes Freight Lines, with terminals in Atlanta, provides intrastate service. Numerous truck lines have interstate authority.

Rail - Southern Railway Co. main line between New Orleans and Washington.

Bus - Greyhound and Southeastern Motor Lines.

Air - Atlanta International Airport (25 miles) is served by 9 airlines.

Utilities:

Electric Power - Douglas County Electric Membership Corp., Georgia Power Co.

Natural Gas - Austell Gas System and Atlanta Gas Light Co.

Water - Douglasville system: Sources - House Creek and small tributary of Little Anawakee Creek. Pumping capacity, 1.3-million gpd; storage capacity, 90-million gallons. Peak demand, 1-million gpd. Douglas County system: Sources - Little Anawakee Creek, capacity of 500,000 gpd; 16- and 10-inch lines from Cobb County; 8-inch lines linked with Villa Rica. Average demand, just over 2-million gpd.

Figure 2. Condensed Facts About Douglasville and Douglas County, Georgia.

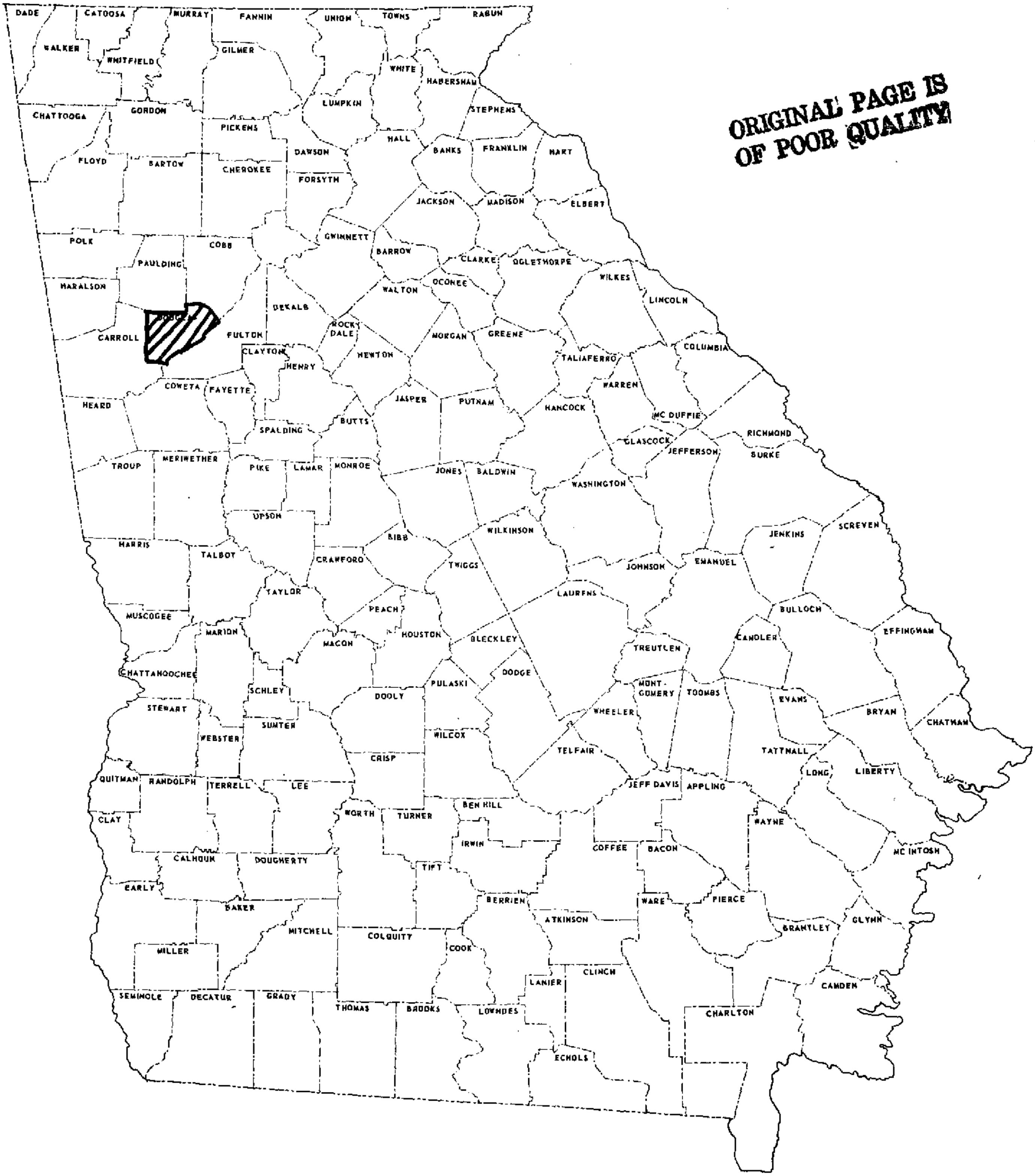

Figure 3. Location of Douglas County.

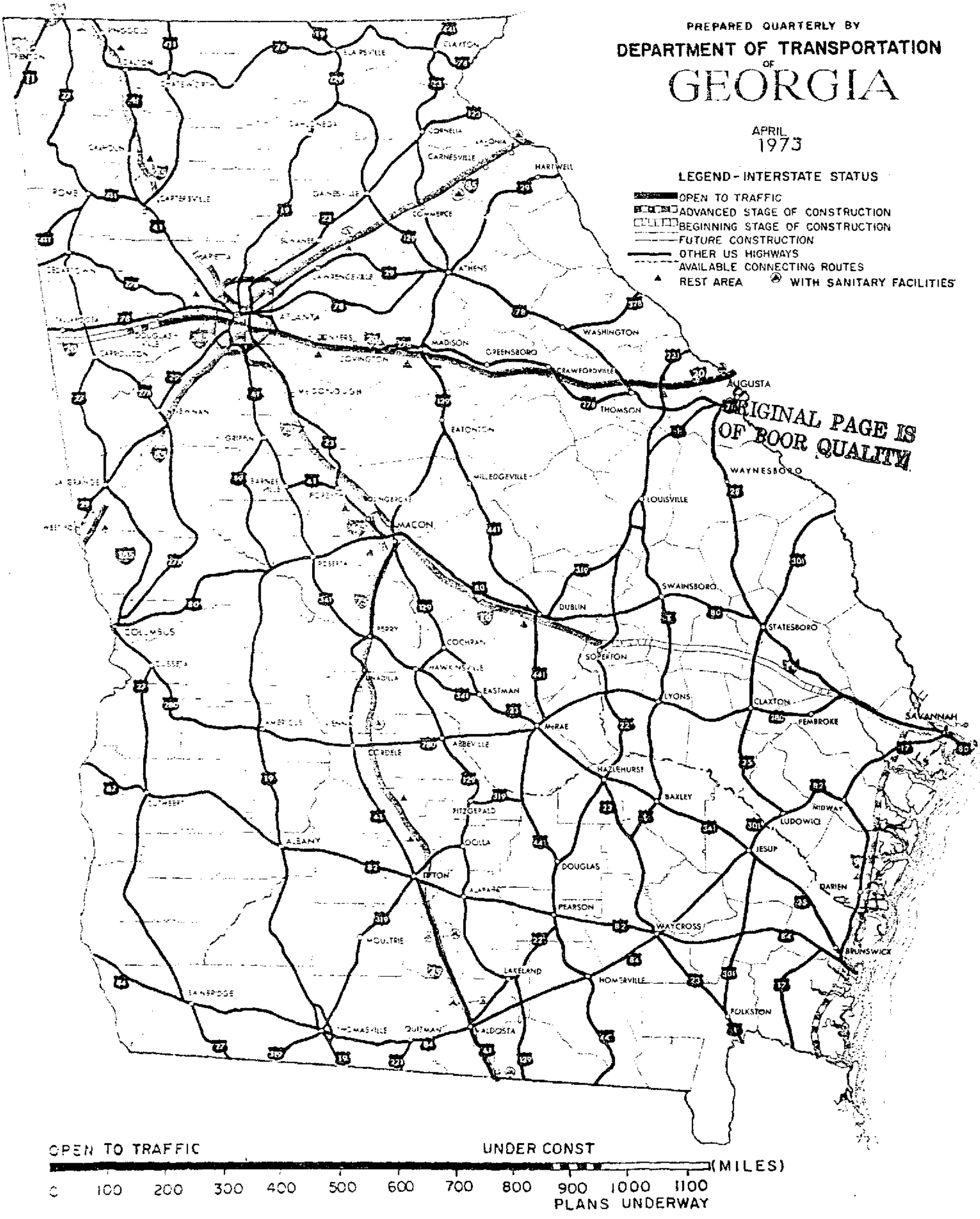

Figure 4. Termination of I-20 in Douglas County.

The present rapid growth and the potential for continued expansion in Douglas County is clearly evident. For the Georgia Department of Natural Resources, then, the results of this study will provide a base of information on the land-use in Douglas County for 1972. It will enable DNR to monitor progress and update this base as appropriate to take into account any of the events mentioned here. If neither of the proposed projects occur, growth within the county will certainly continue, but at a slower rate.

III. DOUGLAS COUNTY LAND-USE MAPPING

Computer Software

The ERTS mapping discussed in this report was accomplished using the Algorithm Simulation Test and Evaluation Program (ASTEP) implemented on a Univac 1108 at Georgia Tech. This program, which was originally written for NASA/JSC has been extensively modified by EES personnel to meet the needs of this and other mapping projects. As currently implemented at Georgia Tech, ASTEP (1) uses a maximum likelihood algorithm for pattern classification, (2) has been modified for automatic scaling specifically for ERTS remote sensing applications, (3) has the capability for rotation of the data to true north and overlaying a geographic coordinate system, and (4) contains provisions for both feature selection based upon a correlation matrix eigenvector transformation and for change-detection pattern recognition.

The maximum likelihood algorithm is based upon Baye's formula* from classical statistics and an assumption of multivariant, normal (Gaussian) probability distributions. (This assumption is usually adequately satisfied in practice, except where multimodal statistics exist.) The algorithm allows supervised classification with greater accuracy than the clustering algorithms if appropriate training data sets are available. Excluding the training time for the classifier, the maximum likelihood approach generally uses less computer time than the clustering method for a specific data set. In addition to the classification algorithm, the program ASTEP contains subroutines which provide the operator with useful statistics, cluster data, and level slices for intelligent use of the program for classification of ERTS remote sensor data. More details on the supervised classification and unsupervised clustering capabilities of ASTEP are contained in Appendices A and B.

Software for operation with a Tektronix Cathode Ray Tube plotter has been integrated into the ASTEP program package. This allows the user to

---

*This allows minimization of the "total expected loss" by individually minimizing the "a posteria conditional risks."

immediately display and generate a hard copy of a 2 or 3 dimensional plot of the spectral data for use in evaluating the separability of data classes. A 2 dimensional histogram of the data may also be selected. By viewing the actual data in 2 or 3 dimensions the user can visually decide if two classes overlap in spectral space. This overlap is often the cause of misclassification.

Land-Use Mapping

Land-use maps have been prepared for that portion of Douglas County which includes Douglasville and the majority of the industrial/commercial/residential land-use in the county. The ERTS scene processed was that of October 15, 1972. NASA high altitude photography, also taken in October 1972, was obtained from the EROS Data Center for use in the accuracy evaluations. Supplemental data in the form of field surveys and low altitude oblique photography were also used.

A "quick look" accuracy evaluation was made to ensure that the land-use categories identified from ERTS were largely correct. This was accomplished by enlarging the high altitude photography to the scale of the ERTS printout - 1:24,000. A visual comparison of the two products then determined that the results were generally correct with the exceptions noted later in this section.

A complete pixel-by-pixel accuracy evaluation is underway. This is being accomplished in the following manner: a clear overlay of the 1:24,000 enlargement is being prepared as a land-use map of the area. Land-use is being classified according to Level II of the USGS/NASA land-use classification system. When complete, approximately 256 square miles will have been mapped. This will be compared with ERTS data of the area to provide quantitative accuracy results for each land-use category. Only partial results will be available for this report. These results are based on supervised classification techniques using maximum likelihood decision criteria.

As stated previously, it is possible to produce land-use maps with a high degree of accuracy using the categories of Level I of the USGS/NASA classification scheme and automatic processing techniques. The categories which can be found and mapped in our test area include: urban and built-up, agricultural land, range land, forest land, water, and barren land. The accuracy of a Level I classification approaches 100%.

The Level II categories which can be identified and mapped include: residential, commercial and services, industrial, extractive, strip and clustered settlement, and open and other; cropland and pasture; deciduous, evergreen, and mixed; streams and waterways, lakes, and reservoir; and bare exposed rock. The categories of Level II present more problems in terms of their unique identification than do the categories in Level I. This is related, in general, to the fact that ERTS measures land cover and we are mapping land-use. These problems, however, will be discussed in more detail later. First we will discuss processing results specifically related to each category above.

Residential. We have been successful in identifying both low and medium density residential as separate categories or as one category. However, we have not found one single category that we could call residential. Multifamily housing, for example, has the same signature as industrial areas in many cases. Hence it could not be completely separated out to be included with residential. There are problems also with identifying heavily wooded subdivisions as residential.

Commercial and Services. Commercial areas, especially those with large parking lots, are readily identifiable. There is good separation between the signatures of commercial and industrial areas. However, there is difficulty in separating commercial and services from institutional which, in fact, often performs some commercial service. An office park does not necessarily look different from an institution of higher learning, for example.

Industrial. The industrial category is reasonably well differentiated from commercial and transportation areas except for transportation/warehousing areas. There are some misclassifications due to large storage areas which resemble manufacturing plants. As was mentioned previously, multifamily housing often has signatures similar to industrial complexes.

Extractive. The only forms of extractive land in the present study area are large stone quarries from which road building materials are derived. These areas are generally identifiable from their high reflectance, but can be confused with concrete parking lots or airport runways.

Strip and Clustered Settlement. This category is identifiable in the processed data but more from its shape than its spectral characteristics. Often this category will contain a combination of commercial, multifamily housing, and transportation.

Open and Other. In an urban/suburban environment this category is most often a well-kept grassy area such as a park, golf course, or cemetary. These areas are identifiable with a high degree of accuracy.

Cropland and Pasture. In the October 15, 1972 scene most of the crops have been harvested. Thus there usually remains only oat or corn stubble, or possibly bare ground where the crops had been planted. Pastures, however, are readily identifiable including some areas which are being grazed after harvesting. The signature for pasture is similar to the open grass areas in more urbanized areas.

Deciduous, Evergreen, and Mixed Forests. Deciduous forests are easily separable from evergreen forests, particularly in October when leaves are turning on deciduous. Mixed forest sometimes tends to be dominated by one category or the other in the classification. However, areas of mixed forest are separable in other instances from either deciduous or evergreen.

Streams and Waterways, Lakes, and Reservoirs. All of these Level II categories tend to be classified into a single category - water. Streams (large) and waterways can be separated from lakes and reservoirs generally on the basis of shape. However, supplementary data are often required to differentiate lakes from reservoirs.

Bare Exposed Rock. No bare exposed rock exists in the areas currently classified in Douglas County. However, from previous studies in the Stone Mountain, Georgia area, it is known that this category can be recognized.

Most of the inaccuracies in classification above relate to trying to classify land-use from land cover. Planners in general, and the Georgia Department of Natural Resources in particular, are interested in land-use information. A heavily wooded residential area with large lots, and hence much space between houses, should be classified as residential from a planner's point of view. However, from the ERTS data it is difficult to classify all of this area into one category which could be called residential. The tendency is to have two or more categories representing forest, grass and housing.

Other examples of this problem are found in the case of airports. One cannot uniquely define an ERTS signature for airports. The area occupied by an airport consists of several different types of land-use including runways and taxiways, buildings, and service/maintenance areas. These and other issues are discussed in more detail in Section VI.

IV. ACCURACY EVALUATION

Preliminary results of our accuracy evaluation of the computer generated land-use map are given in this section. For the purposes of this report only about 10% of the total area was evaluated. Hence, these results are subject to change when a more complete evaluation is made.

The photointerpretation was assumed to be correct. Both NASA high altitude photographs and low altitude observations and field checks were used in arriving at the "correct" classifications. However, the results may be subject to some revision as the study proceeds.

The overall accuracy of the computer-generated map was 67% as shown in Table I. Accuracies ranged from 87% in the residential category to only 26% for the open category. This low figure results, in part, from an inadequate sample containing open areas and the diverse definition given to open areas.

An area of substantial misclassification was in the three forest categories—deciduous, evergreen, and mixed. Had there been only one category into which all forest areas were classified, the overall accuracy would have risen to 79%. Land-use maps generated by and for planning agencies typically have only one category for forest, and this may be a transparent color overlaying all other categories.

While this accuracy is certainly not as high as is desired for most land-use maps, the results compare favorably with published results of manual photointerpretation of high altitude photography. In a recent report by Paul L. Vegas [6] at NASA/MSTL, an overall accuracy of 84% was obtained using manual interpretation of NASA high altitude photography. The categories used in the classification were somewhat different than those for Level II categories. However, there is enough similarity to warrant comparison. The results of this test are displayed in Table II.

Most of the area (approximately 95%) of Georgia is rural. Since the accuracy of this technique is highest in rural areas, it is estimated that 95% of the area of Georgia could be mapped with accuracies in the 80% to 90% range.

	Res.	Com.	Ind.	Extr.	Trans.	Open	Crops	Decid.	Ever.	Mixed	Water	% Accuracy
Residential	1056	29	5	0	36	2	39	7	30	11	0	87
Commercial	67	178	36	2	11	6	6		2			58
Industrial												-
Extractive		7	3	16								62
Transportation	56	2			43							43
Open	17				3	7						26
Crops	50						105	2	42	3		52
Deciduous	70				2	1		298	34	145		54
Evergreen	45				1			7	190	53		64
Mixed	126	1						57	88	401		60
Water	1				1		1	1			14	78
TOTAL												67
TOTAL (with only 1 forest category)												79

TABLE I. Accuracy of Computer Generated Land-Use Map from ERTS Data. (Numbers in Matrix Indicate Number of ERTS Pixels.)

		R	C	I	P	ROW	W	M	G	F	Cu	H	G	O	% Accuracy
R Residential	48	45	2							1					94+
C Commercial	30	14	16												53+
I Industrial	11			11											100
P Public, Public/Semi	21	6	2		12				1						57+
ROW Right of Way	27					27									100
W Water	30						30								100
M Marsh	24						1	17	2	4					70+
G Grassland	42							1	24	2	1				86
F Forests	59							1		58					98+
Cu Cultivated	29							2	6		20			1	68+
H Horticulture	14									1		13			92+
O Other	1													1	100
Total	336											Overall Average			84+

TABLE II. Accuracy of Land-Use Classification by Photo Interpreter. (Numbers in Matrix Indicate Number of Sample Points.)

V. OTHER RESULTS FROM ERTS PROCESSING

The land-use analysis of Douglas County has brought about an unexpected geological result that may be extremely important if extended to other areas. In the process of obtaining training for supervised land-use classification of the Atlanta and Douglas County areas, unsupervised classification computer runs were made to isolate clusters or to separate things that "looked" different from one another. In the Atlanta Cumberland Mall area two basic groups of trees were identified and were used as training classes for the Douglas County area. When supervised classification was made, a definite elongation was noticed in the distribution of the second type of trees. This North Eastward elongation was north of the Chattahoochee River and was parallel to the river. When this trend was traced back to the east toward Atlanta, it was found that even though the river bent sharply to the north, the trend remained parallel to the direction of the river in Douglas County. In fact, the elongation seems to parallel the Brevard Shear Zone (Reference 6), a major geologic trend, rather than the river itself.

Through the assistance of Mr. J. F. Brooks of the Soil Conservation Service, a soils map was obtained for Douglas County [7]. When this map was analyzed, a major soil zone was found to lie parallel to the Chattahoochee and approximately in the same position as the elongated tree zone detected in the ERTS data. The soil group (Louisa Fine Sandy Loam) consisted of excessively drained strongly acidic soils formed in materials weathered from mica schist. These soils have slopes ranging from 10 to 40 percent with two thirds of the acreage between 15 and 25 percent. The supply of organic matter is medium to low in the Louisa soils. Thus, the soils are poorly suited to crops or pasture, but are well suited to loblolly and shortleaf pines and to plants that provide shelter and food for wildlife. More than 95% of the acreage is in trees with varying degrees of slope. By overlaying a scaled computer output on the soil map an excellent correlation was found between the soil type and the ERTS vegetation tree type. The area discussed above is intensely forested so the ERTS data could not be differentiating soil zones directly; however, the computer analysis of ERTS data does detect a change in the vegetation types indicative of different soil types. This type of indirect information is often as useful as direct information.

Future analysis should be designed to trace this trend further to the east and west and thus to prove or disprove the hypothesis that the ERTS data may be detecting a shear zone by vegetation differences.

VI. PROBLEMS RELATIVE TO ERTS PROCESSING USING USGS/NASA LAND-USE CLASSIFICATION SYSTEM

Introduction

A project review meeting held at MSFC in August, 1974, provided a forum for discussing some of the problems associated with computer generated ERTS land-use maps. Those in attendance at the meeting discussed ERTS computer land-use mapping from a general standpoint and also with specific reference to the present project. Many of the issues raised in this section result from comments made in this project review.

Some categories of land-use are not obtainable from any remote sensor - ERTS or high or low altitude photography. Consider the categories of transportation, communications and utilities. From ERTS or from photography, an airport will not look similar to a rail switching yard, let alone a communications complex or a utility. A human interpreter can possibly make allowances because of *a priori* knowledge and classify all of the above into a single category. However, it is not possible for even a human interpreter to exactly define the boundaries of the above unless they are fenced in at the boundary or there is a change of vegetation at the boundary.

Many other categories share this problem. It can be difficult to discern the boundary of a park, for example, from either photographs of ERTS computer maps. Clearly supplemental information is required to make a land-use map which accurately reflects parameters necessary for intelligent planning.

Part of the problem with an airport, for example, is that there are several types of land cover within the boundary. At the Hartsfield International Airport in Atlanta, there are these categories of land cover: bare ground, concrete, asphalt, large buildings, trees, and grass. On a computer classification map these areas are likely to classify with industrial, commercial, forest, and open and other.

The preceding paragraph outlines a problem which is much more general than just defining the boundaries of a particular category such as transportation/airport. This is the problem of observing land cover and classifying land-use. It is apparent in several categories of land-use. Residential areas, for example, range from apartment complexes to cluster/condomium homes to single family detached residences with lot sizes from 1/4 acre to 10-15 acres - even in urban areas. It appears that planners generally would like for all of these to be categorized as residential or possibly multifamily/single family residential.

This has proved impossible so far. The difficulties with multifamily have been discussed previously. Contextual information (or *a priori* knowledge) however, often allows one to differentiate between industrial areas and multifamily residences. With very low density residential areas, particularly those which are heavily wooded, there are likely to be several categories on a computer generated ERTS map. The areas occupied by the houses/lawns/driveways will probably be classified in a category which includes higher density single family residential. The forested areas in between houses, however, are likely to classify as deciduous, evergreen, or mixed. Since these areas are neither open/other nor forests in the true sense of the word, they should be classified residential. (Indeed there is no category for forest in class 01.) This has proved difficult so far, because to classify these areas accurately would require a decision algorithm incorporating spatial/contextual information.

Another problem arises in a test area such as ours which includes both urban and rural land-use. Open areas in an urban setting are usually golf courses, parks or other grassy areas. The signature for this category of land-use is virtually identical to the signature for pastures - a rural land-use. While each of these categories can be identified in its proper setting, there are no unique signatures which apply to these categories separately.

There are other problems associated with measuring land cover and mapping land use but these generally are similar to the above. It seems that two additional questions need to be addressed in order to cope with these problems.

(1) What is the minimum complement of additional information that will enable one to produce accurate land-use maps?

(2) What additional processing techniques are available to provide some of the spatial and contextual information required?

These two questions are discussed in more detail below. Some of the processing techniques discussed will be given a preliminary examination during the remainder of the project.

Supplemental Mapping Information

The most logical place to start looking for additional information is on USGS 7 1/2 minute quadrangle maps. These maps suffer from infrequent updating and incomplete coverage, but this need not be a severe handicap. Some of the more difficult categories of land-use are semi-permanent--transportation facilities, for example. Other useful information of a semi-permanent nature is also available including parks, schools, churchs, cemetaries, hospitals, prisons, etc. One could start the mapping project with these land-uses on a base map and concentrate the ERTS data processing on other categories such as residential, commercial and industrial. These are the categories that change rapidly - particularly in a fast-growing urban/suburban area. In contrast, the boundaries of parks, airports, etc., change slowly, if at all, and these boundaries are shown on the USGS maps.

Another source of useful information is visual examination of the area. The traditional windshield survey, however, is quite slow and tedious. A more efficient method for these examinations seems to be low altitude surveillance from light aircraft. In our current project the two investigators spent a major portion of one day visiting approximately two dozen sites in Douglas County and photographing these areas. A return visit was made by light aircraft and the same sites, plus many others, were photographed in less than 1 hour flying time and less than two hours total time.

The above are some possible sources of supplemental information which would be useful to an ERTS computer mapping project. In those operational cases where they are employed, there seems to be no system for carrying out these tasks in an efficient and timely manner. It seems, therefore, that work to devise and test such a system would be beneficial to those who require land-use information on a regular basis.

Proposed Techniques for Additional Processing

In addition to the supervised and non-supervised classification techniques already in use in this project, several other techniques are proposed as possible methods of extracting more information out of the ERTS than is currently available. Some of these methods will be used singly while others may be used as supplemental spectral information.

Ratio processing has been used extensively in the analysis of multispectral data. Reference 8 is an example of studies using this technique for various applications. Two channels of data may be ratioed as a normalization procedure which should eliminate any brightness variations within an ERTS scene. This ratioed data may be analyzed separately with a level slicing technique or may be added as a fifth ERTS channel of data. A simple data reduction technique might be to ratio all channels of data to channel one of ERTS data and classify only on the 3 ratioed bands. An investigation should be made as to the usefulness of such a technique for land-use applications.

Linear decision theory should be considered as a rapid method for classification of large ERTS data sets for land-use information. For a regional study the loss in accuracy from that of a quadratic technique might be an acceptable tradeoff with the computer time needed to produce the desired result. The use of linear decision theory as a tool for analysis of MSS data is depicted in Reference 9.

The methods so far considered in this study have contained no mechanism for the inclusion of spatial information in the land-use classification from ERTS. The inclusion of spatial characteristics in the ERTS classification provides an extra source of information that may prove valuable in land-use and other studies. This spatial data provides information on the texture of various subsets of ERTS data. Lineation detection is one use of the spatial information and may have profound uses in geological investigations. The spatial information described above is provided by performing a fast Fourier transform (FFT) on an n x n subset of ERTS data in one channel where n is the number of pixels considered (n = a power of 2). The FFT technique is described in Reference 10. The FFT results may be utilized in several different ways. If one is looking for lineations that may represent highways, faults, etc., the log magnitude of the transform may be displayed. Lineations in the picture before the FFT are shown as lines through the center of the star diagram in the same direction as the original lines were relative to the picture.

The use of the FFT could provide an efficient method for the data compression of ERTS data since the FFT is inherently symmetric. The derivation of a distance measure in the transformed space would allow classification in FFT space instead of spectral space. This technique would allow inclusion of spatial data into the actual classification scheme.

An alternative method is to use the integral of the FFT over 3 different regions as additional channels of data to be used in classification. Reference 11 indicates success in recognizing various types of physical morphology by using this type of process. Tests should be made to see if this method would benefit land-use classification.

One of the major problems encountered with ERTS data is that the minimum size of one pixel is approximately one acre. The reflectance received at the spacecraft is normally not that from a uniform substance, but instead may be from a mixture of several different things that occupy that acre of ground on the earth. For example, a pixel in a subdivision may actually consist of reflectances from houses, grass, trees, asphalt, and concrete. This problem is accentuated when one has to choose training classes for particular land-use categories. In the early analysis of multispectral scanner data, the data were obtained from low flying aircraft. The pixel size therefore was very small compared to that of ERTS. Training classes were chosen by using ground truth to identify areas that were covered by a certain crop. The MSS data over this area were aggregated and a statistical analysis was performed. In most cases, the data in each channel were normally distributed about a mean value for a particular training field. If the data were unimodal, the statistics provided a means for identifying other pixels that might contain the same crop. This was called Maximum Likelihood Classification. The fact that we should not lose sight of is that even though the field was unimodal, the variation about the mean was caused by a mixture of the reflectance from the crop and the reflectance from the surrounding soil. Thus, even in the earliest days the mixture problem was with us.

By increasing the size of the resolution element to one acre, we are not treating a different problem. Most researchers tend to ignore the mixture problem in the hopes that it will fade away; however, it will always remain with us. Reference 12 describes a mixture analysis scheme that was developed by TRW Systems using the ASTEP program. The current version of ASTEP in operation at Georgia Tech has a similar mixture algorithm incorporated into it. Basically, the method assumes a pixel is a linear combination of several "pure" signatures, i.e., grass, pines, water. By applying the mixture technique the proportion of each pixel covered by each of the signatures may be estimated. This, of course, assumes: (1) pure signatures can be generated, (2) they are linearly additive, and (3) all elements of the mixture are known. The adaptation of a mixture algorithm into a land use study would normally include phases for testing of the algorithm, inclusion of the mixture algorithm into an efficient classification system, and an evaluation of the aesthetics involved in defining land-use rather than land cover categories.

References

1. Ellefsen, Richard, Philip Swain, and James Wray, "Urban Land-Use Mapping by Machine Processing of ERTS-1 Multispectral Data: A San Francisco Bay Area Example," Machine Processing of Remotely Sensed Data, Purdue University, October 16-18, 1973.

2. Todd, William J. and Marion F. Baumgartner, "Land-Use Classification of Marion County, Indiana by Spectral Analysis of Digitized Satellite Data," Machine Processing of Remotely Sensed Data, Purdue University, October 16-18, 1973.

3. Raje, S., R. Economy, and J. McKnight, "'First Look' Analyses of Five Cycles of ERTS-1 Imagery Over County of Los Angeles: Assessment of Data Utility for Urban Development and Regional Planning," Symposium on Significant Results Obtained from the Earth Resources Technology Satellite-1, NASA SP-327, pp. 1023-1030, March 5-9, 1973.

4. Todd, William, Paul Mausel, and Kenneth Wenner, "Preparation of Urban Land Use Inventories by Machine Processing of ERTS MSS Data," Symposium on Significant Results Obtained from the Earth Resources Technology Satellite-1, March 5-9, 1973.

5. Anderson, James R., Ernest E. Hardy, and John T. Roach, "A Land-Use Classification System for Use with Remote-Sensor Data," Geological Survey Circular 671, U. S. Geological Survey 1972, Washington, D. C. 20242.

6. Crawford, T. J., and Medlin, Jack H., "Brevard Fault Zone in Western Georgia and Eastern Alabama," Guidebook 12, Geological Society of America, Southeastern Section Annual Meeting, 1974.

7. Wells, R. D., McIntyre, C. L. and Ritchie, F. T., Jr., "Soil Survey of Douglas County, Georgia," United States Dept. of Agriculture, Series 1959, No. 1, November 1961.

8. Dillman R. and Vincent, R., "Unsupervised Mapping of Geologic Features and Soils in California," Ninth International Symposium on Remote Sensing of Environment, Ann Arbor, Michigan, April 15-19, 1974.

9. Bond, A. D., and Atkinson, R. C., "An Integrated Feature Selection and Supervised Learning Scheme for Fast Computer Classification of MSS Data," Conference on Earth Resources Observations and Information Analysis Systems, Tullahoma, Tennessee, March 13, 1972.

10. Cochran, et al., "What is the Fast Fourier Transform?" Proceedings of IEEE, Vol. 55, No. 10, October 1967.

11. Gramenopoulos, Nicholas, "Finding the Earth in ERTS," Optical Spectra, p. 26, August 1974.

12. Salvito, P., Jr., "Iterative Techniques to Estimate Signature Vectors for Mixture Processing of Multispectral Data," Conference on Machine Processing of Remote Sensed Data, Purdue Univ., October 16-18, 1973.

APPENDIX A

UNSUPERVISED CLASSIFICATION OF ERTS MSS DATA

As discussed in Section 2 of this paper, each resolution element for the Earth Resources Technology Satellite scanner system represents an area on the ground of approximately 1.05 acres. Each resolution element in turn has a set of four measurements associated with it. These four measurements are the intensities of light received by the detectors on the spacecraft in each of four spectral bands and may be considered a four-dimensional vector associated with each plot of ground. We would like to have some intuitive feeling for where the tip of this vector is located in four space. Unfortunately, four dimensions is difficult to visualize, so for an example we will take a three-dimensional vector. This might represent measurement in three regions of the spectrum instead of four. Now if we let each axis of a coordinate system represent intensities in one spectral region, we can visualize the location of each vector in three space. For example, let us have three measurements (Vector A) normalized between 0 and 256:

Reading	Axis	Spectral Region
222	x	.5 - .6 microns
250	y	.6 - .7 microns
210	z	.7 - .8 microns

Figure A-1 shows the location of this vector in three space. If we have another data vector B associated with a different area:

Reading	Axis	Spectral Region
234	x	.5 - .6 microns
220	y	.6 - .7 microns
230	z	.7 - .8 microns

Figure A-2 shows the location of vectors A and B in three space. Now, we would like to have some measure of the difference between measurement vector A and B. The most logical choice for a difference measure is the distance between the two vector tips. This distance is given by

$$d = |\bar{A} - \bar{B}|$$

where $| \ |$ indicates absolute value or a magnitude of a vector and $\bar{A}$ means

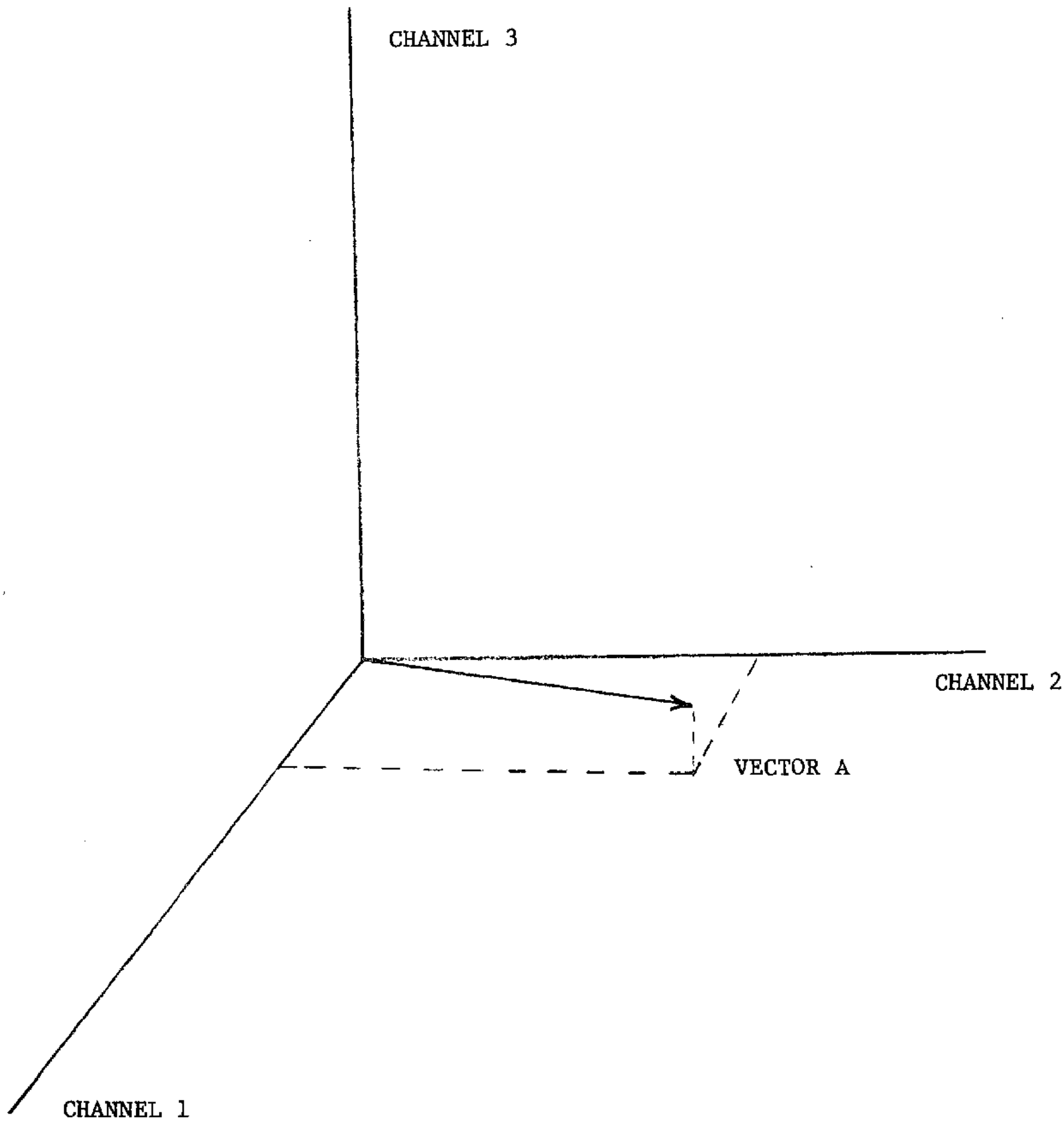

Figure (A-1)

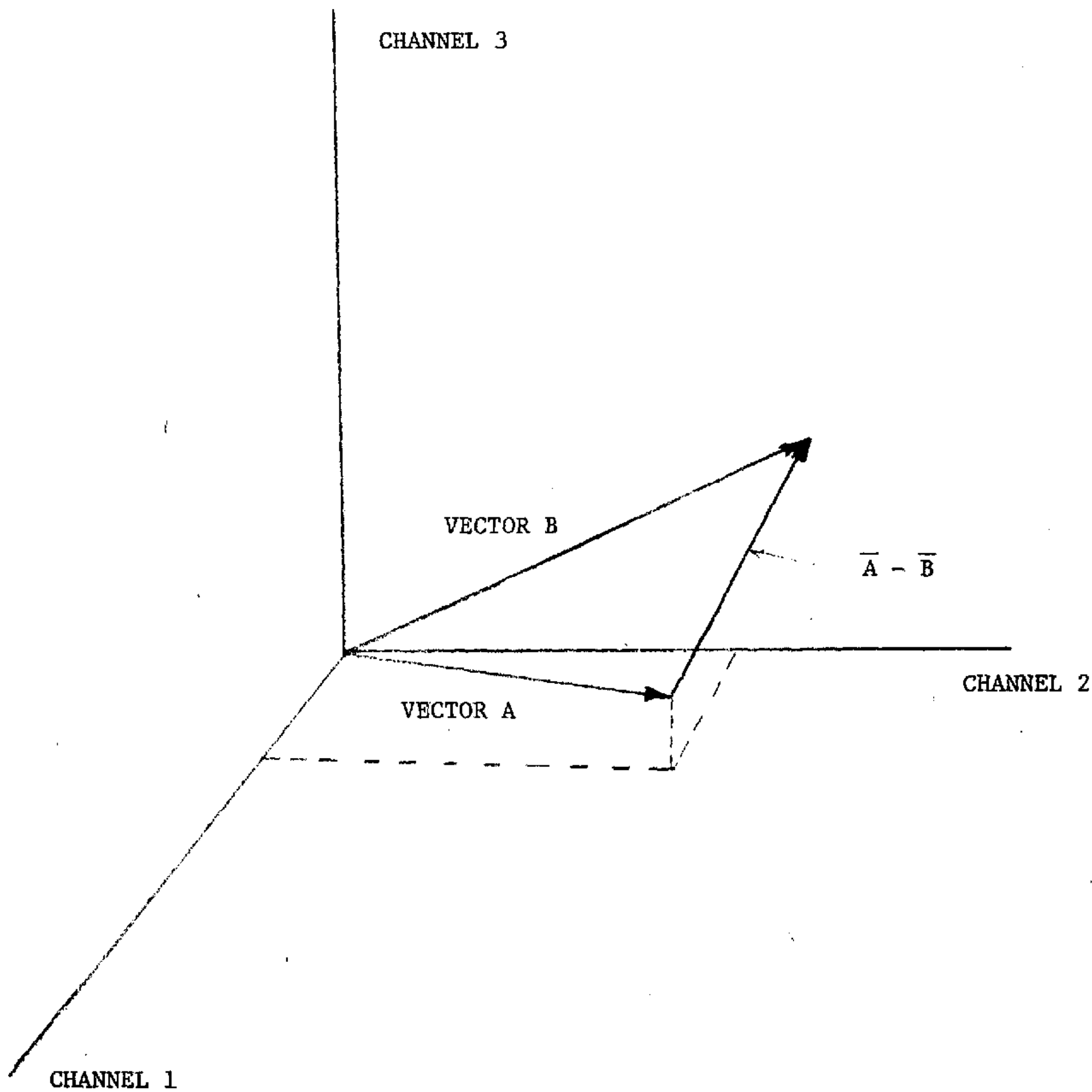
CHANNEL 3
VECTOR B
$\overline{A} - \overline{B}$
CHANNEL 2
VECTOR A
CHANNEL 1

Figure (A-2)

that A is a vector quantity. Expanding to evaluate d, we have

$$d = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2}$$

where a_1 and b_1 are the first components of the vectors $\overline{A}$ and $\overline{B}$. The angle between A and B may also be calculated by

$$\Theta = \cos^{-1} \left(\frac{\overline{A} \cdot \overline{B}}{|\overline{A}| \ |\overline{B}|}\right)$$

where A · B is the inner product of A and B; i.e.,

$$A \cdot B = a_1b_1 + a_2b_2 + a_3b_3$$

Therefore,

$$\Theta = \cos^{-1} \left(\frac{(a_1b_1 + a_2b_2 + a_3b_3)}{(a_1^2 + a_2^2 + a_3^2)^{1/2} \ (b_1^2 + b_2^2 + b_3^2)^{1/2}}\right)$$

in terms of components of $\overline{A}$ and $\overline{B}$.

It can be seen that in four dimensions

$$d = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2 + (a_4 - b_4)^2}$$

and

$$\Theta = \cos^{-1} \left(\frac{(a_1b_1 + a_2b_2 + a_4b_4)}{(a_1^2 + a_2^2 + a_3^2 + a_4^2)^{1/2} \ (b_1^2 + b_2^2 + b_3^2 + b_4^2)^{1/2}}\right)$$

These equations will be used later. Another quantity that we would like to define is the mean vector. This vector is essentially the average vector associated with a set of N vectors. It is calculated by

$$\overline{M} = \frac{1}{N} \sum_{i+1}^{N} \overline{A}_i$$

where A_i is the i^{th} individual vector. In terms of four components, we have

$$M_1 = \frac{1}{N} \sum_{i=1}^{N} (a_1)_i = \frac{1}{N} (a_{11} + a_{12} + a_{13} + \ldots + a_{1N})$$

$$M_2 = \frac{1}{N} \sum_{i=1}^{N} (a_2)_i = \frac{1}{N} (a_{21} + a_{22} + a_{23} + \ldots + a_{2N})$$

$$\vdots$$

$$M_4 = \frac{1}{N} \sum_{i=1}^{N} (a_4)_i = \frac{1}{N} (a_{41} + a_{42} + a_{43} + \ldots + a_{4N})$$

where a_{21} is the second component of the first vector considered and a_{23} is the second component of the third vector considered.

Now consider the situation in Figure A-3. The multispectral scanner scans a region normal to the flight path of the spacecraft. At any instant in time the rotating mirror displays an image representing approximately one acre on the ground and measurements in 4 regions of the spectrum are taken. The spacecraft velocity and the scanner rotation speed are such that after one scan line of data is taken, the spacecraft has moved forward enough so that the next scan line is contiguous to the first.

The massive amount of data that is taken for one ERTS scene of 100 nm x 100 nm can be analyzed digitally using unsupervised classification and the quantities described above. Each resolution element's radiance values are represented in four-space, and we would like to decide which resolution elements resemble others in an ERTS scene. A typical situation in three-space is shown in Figure A-4. It can be seen that there are several groupings of data points which probably represent radiance values from the same or similar objects. For example, group A might be radiance values from trees, Group B from buildings, and Group C from water. Using the techniques developed above we may crudely represent each group or cluster by a mean vector and a chosen radius in three-space (Figure A-5). Any radiance vector that falls within this radius of the Group A mean is assigned to Group A. This follows similarly with other groups. If a vector does not fall within the prescribed radius

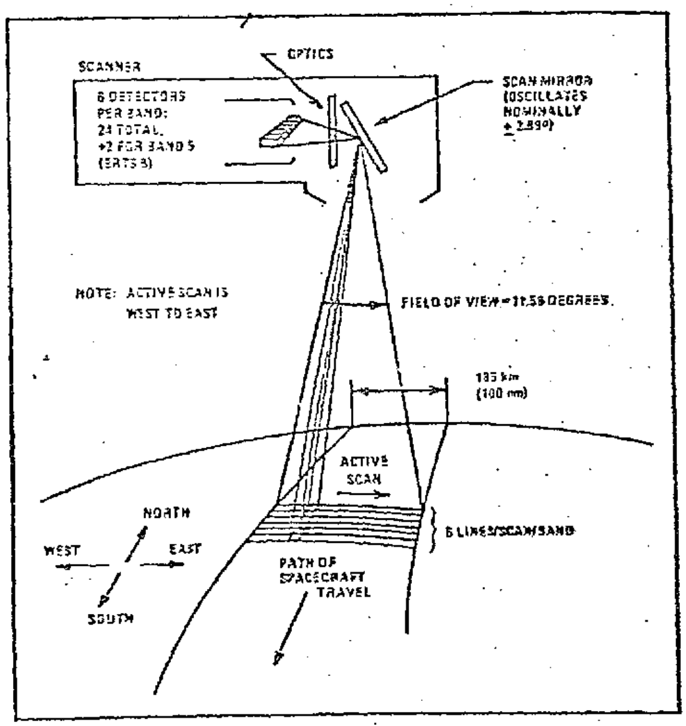
SCANNER
OPTICS
6 DETECTORS
PER BAND:
24 TOTAL
+2 FOR BAND 5
(ERTS B)
SCAN MIRROR
(OSCILLATES
NOMINALLY
± 2.89°)
NOTE: ACTIVE SCAN IS
WEST TO EAST
FIELD OF VIEW = 11.56 DEGREES.
185 km
(100 nm)
ACTIVE
SCAN
6 LINES/SCAN/BAND
NORTH
WEST
EAST
SOUTH
PATH OF
SPACECRAFT
TRAVEL

Figure (A-3)

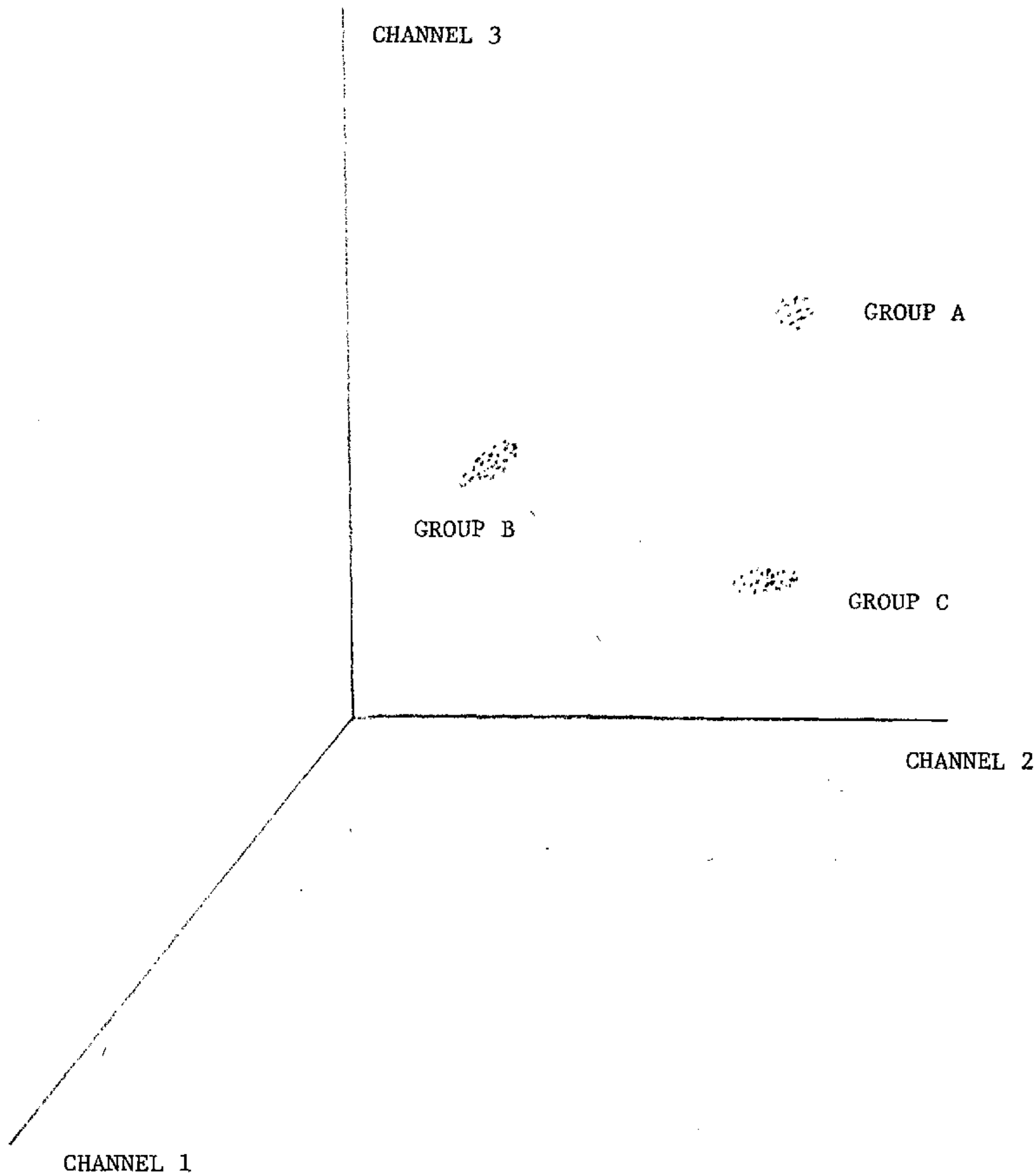

Figure (A-4)

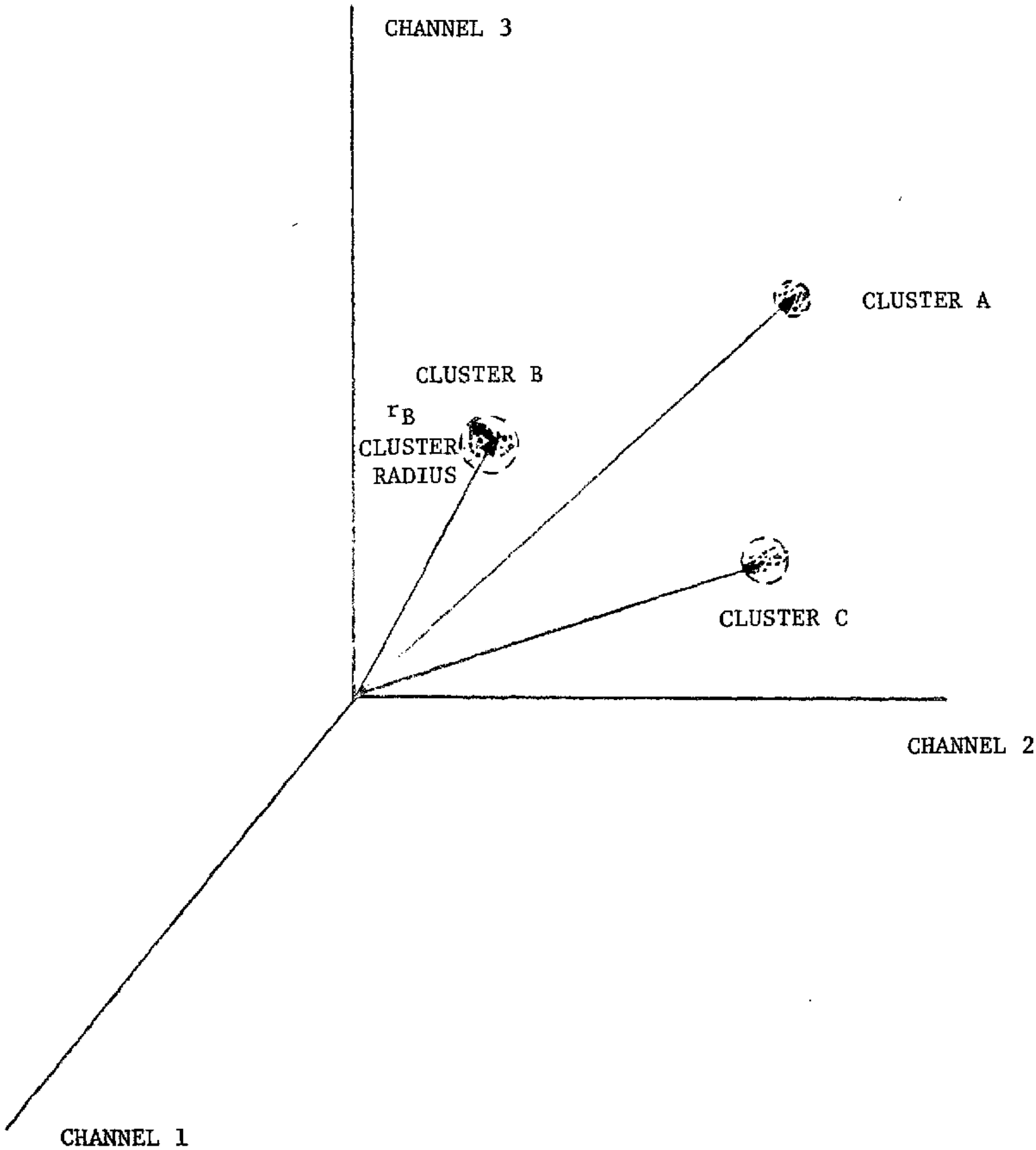
CHANNEL 3
CLUSTER A
CLUSTER B
r_B
CLUSTER
RADIUS
CLUSTER C
CHANNEL 2
CHANNEL 1

Figure (A-5)

of any of the previously defined clusters, a new cluster is generated using that vector as the first point. The data is usually considered sequentially considering one resolution element at a time for a whole area. One obvious disadvantage is that if the radii are chosen too small only a few points are allowed in a cluster and many additional clusters will have to be formed. The selection of radius values is essentially a trial and error procedure. As the number of clusters increases so does computer time and storage. This limits the number of clusters that may be considered. The present limit for our computer program is 20 clusters. If the program determines that a 21st cluster should be formed, then a statistical method considering the number of points in each cluster is used to decide which of the original clusters to eliminate. Actually a user of the program may set the maximum number of clusters to any number he likes up to 20.

The ASTEP (Algorithm Simulation Test and Evaluation Program) utilizes a sequential clustering as described above with minor modifications. Two iterations are made through the entire data set. The first iteration considers each measurement vector separately; i.e., the first vector is the first cluster; the second vector, if it is not within the specified radius of the first cluster, forms a second cluster and so on. If it is, the two vectors are averaged to form the cluster mean. It can be seen that this method may be biased due to the starting point in the data set. To eliminate this bias, a second iteration is made not allowing the mean vectors to be updated sequentially. The final product is a set of less than 20 groups of objects or things that look similar. These groups may often be associated with different objects on the ground such as water, rock, etc. These programs require a great deal of experience to determine radius values that will separate natural objects on the ground. A computer printout may be generated that represents the area that the satellite has imaged. Each character on the printout is associated with one of the previously determined clusters. Thus one can see the spatial location of similar and dissimilar things on the earth's surface. With some checking with maps and aerial photos, these clusters may be used to represent major housing and development trends within a city as well as many other uses including geological.

APPENDIX B

SUPERVISED CLASSIFICATION OF ERTS MSS DATA

ERTS supervised classification is different from unsupervised classification in that instead of having a digital technique find separate clusters of measurement vectors in four-space, a method is asked to classify each measurement vector into one of several classes whose position in four-space has been previously computed. Each class in the supervised method represents a particular physical characteristic of the area imaged by the ERTS multispectral scanner system. For example, supervised classes may be defined as water bodies, commercial areas, cleared land, etc. To completely define a class we need more information than was used in unsupervised classification. Instead of a mean vector and a radius around it describing a class, we now use a method which allows us to describe the shape of the envelope surrounding all points in one class. For example, in clustering we assumed that the points were symmetrical about the mean vector. Much statistical work has been done that indicates that most natural phenomena may be adequately described by a mean vector with a normal distribution of points around it, and not by a mean vector with an envelope equidistant in all spectral channels. In three-space a normal distribution resembles an ellipsoid about the mean (Figure B-1). Thus, if we wanted to describe an ellipsoid in three-space we would need to calculate the mean and the direction and length of the semi-minor and major axes. This may be done in three-space and extended into n-space by the calculation of the varience of the data from the mean. The variance denoted by σ^2 is a measure of the elongation of the data in a particular direction. It may be calculated by standard statistical methods. An intuitive feeling for σ is found by the following equation. In 95% of the cases considered a random data value x will fall in the region defined by $|x - \mu| \leq 2\sigma$ where μ is the mean value. Figure B-2 shows the region for one dimension. σ may be considered to be a difference in spectral response in one channel from the mean value. This may be extended to N channels of data by considering that there is a variance associated with each channel of data. Since we are dealing with data randomly distributed within a normal distribution, we can only estimate the values for the mean and the variance

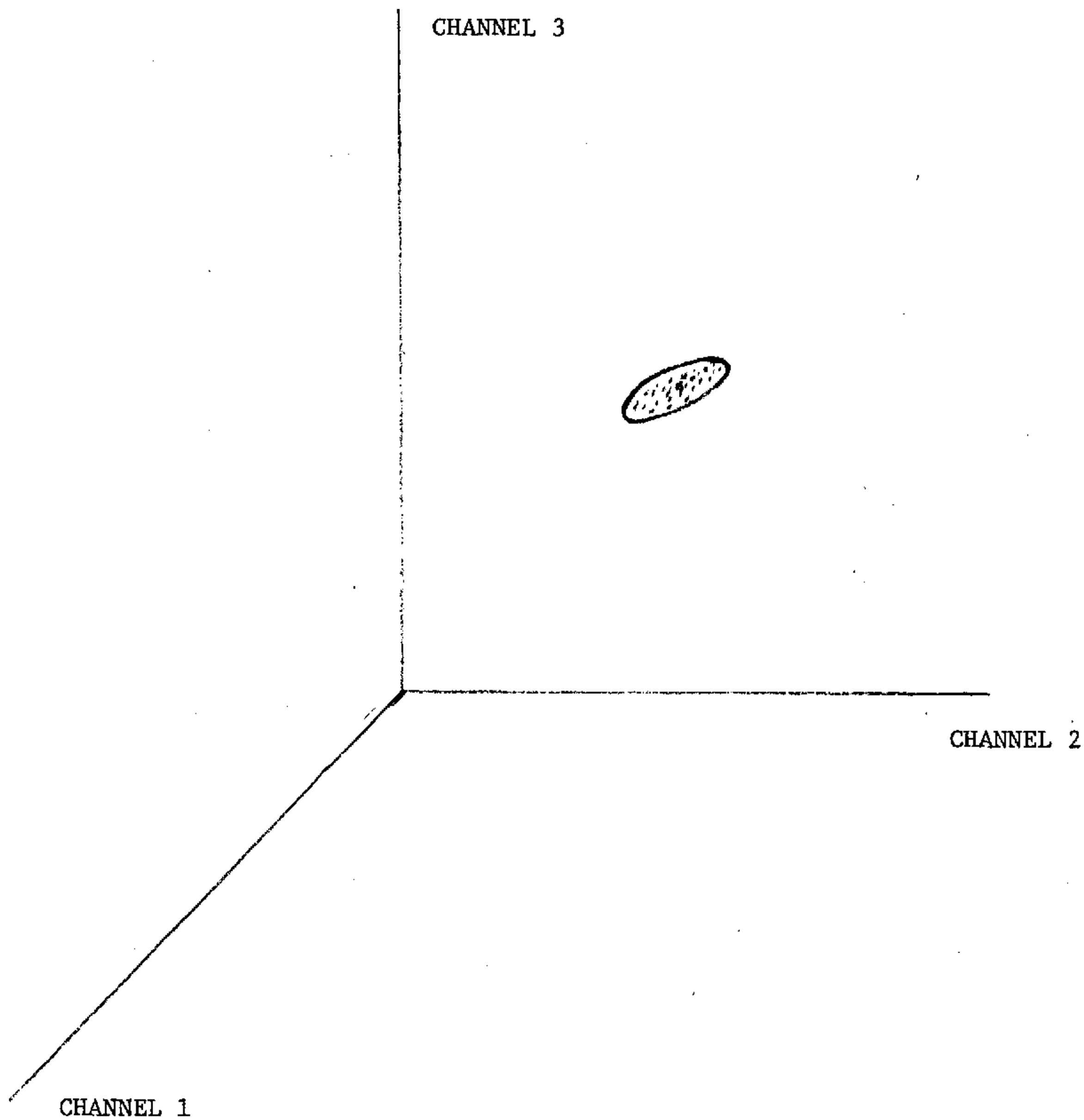

Figure (B-1)

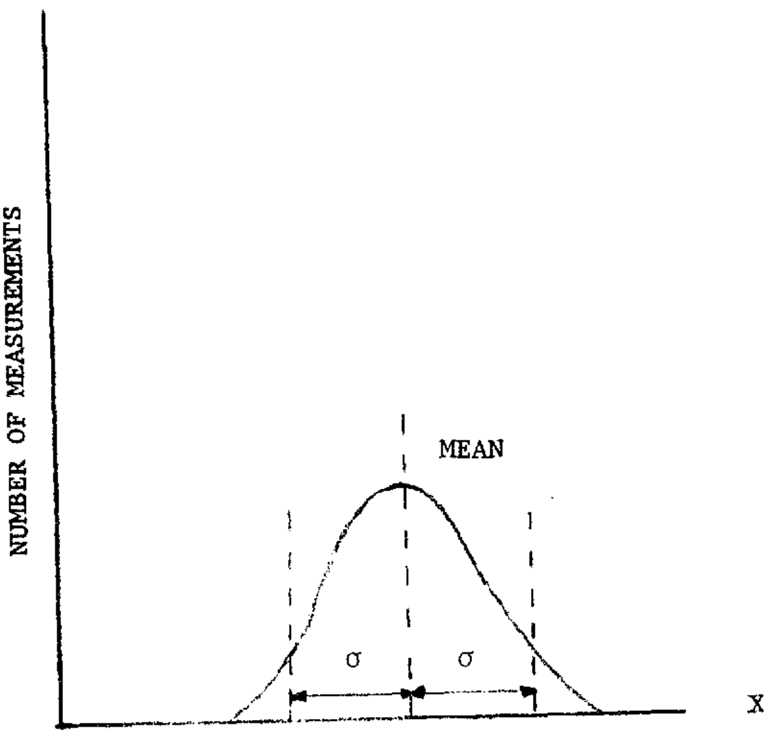

Figure (B-2)

associated with a particular class. In general, if a large number of samples are considered to calculate the mean value, the mean will approximate the true mean. If only a small number is considered there may be significant error in the calculation of the mean for a particular class. In multivariate analysis, the variances in each of the spectral regions are not the only considerations. If data values in some channels depend on data values in other channels, there will be a covariance between the two channels of data. For N channels this may be represented in an N by N matrix (the covariance matrix). If there is no interdependence, the channels are said to be independent and the covariance is zero. The best estimate for the mean and covariance matrix is given below.

$$\hat{\mu} = \frac{1}{N} \sum_{k=1}^{N} \overline{X}_k \qquad \text{where X is a single data vector}$$

and

$$\hat{\Sigma} = \frac{1}{N} \sum_{k=1}^{N} (\overline{X}_k - \hat{\mu})(\overline{X}_k - \hat{\mu})^t$$

where the t indicates the second matrix is transposed. If a sufficient number of samples are used to define the above population, the diagonal elements of the covariance matrix will be the variances squared for each channel and the off diagonal elements describe the interreaction between channels of data. A sample case for 3 channels is shown below.

$$\hat{\Sigma} = \begin{matrix} \sigma_1^2 & \sigma_1\sigma_2 & \sigma_1\sigma_3 \\ \sigma_2\sigma_1 & \sigma_2^2 & \sigma_2\sigma_3 \\ \sigma_3\sigma_1 & \sigma_3\sigma_2 & \sigma_3^2 \end{matrix}$$

If the channels of data were independent then

$$\hat{\Sigma} = \begin{matrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{matrix}$$

Thus given a sufficient number of radiance vectors that are identifiable with one class of natural phenomena, an estimated mean and a covariance may be computed for the total population of that phenomena. By comparing each data vector to these estimates, we may decide if that data vector in fact represents a certain class of material, i.e., water. This will be discussed further below.

Discriminant functions are developed in classification theory for special distributions of data. These discriminant functions are the criteria by which a radiance vector may be assigned to a particular class. Since the normal density function is very often used to represent reality, the discriminant function for it has been known for some time. The discriminant function for a radiance vector $\overline{X}$ to be in the i^{th} class is

$$g_i(\overline{X}) = -1/2 \; (\overline{X} - \overline{\mu})^t \; \Sigma_i^{-1} \; (\overline{X} - \overline{\mu}) - \frac{d}{2} \log 2\pi - 1/2 \log |\Sigma_i| + \log P(w_i)$$

where $\overline{\mu}$ is the mean vector and Σ_i^{-1} is the inverse of the i^{th} class covariance matrix. In general the $\frac{d}{2} \log 2\pi$ term is only additive and is not a function of which class is considered. Thus it may be ignored. By replacing $g_i(X)$ by $f(g_i(\overline{X}))$ where f is a monotonically increasing function, the resulting classification is unchanged (Ref. 1). Thus if we take the exponential of $g_i(\overline{X})$

$$Q_i = f(g_i(\overline{X})) = \frac{e^{-1/2(\overline{X} - \overline{\mu})^t \; \Sigma_i^{-1} \; (\overline{X} - \overline{\mu})}}{|\Sigma_i|^{1/2}}$$

Now for every radiance vector $\overline{X}$ a Q is calculated for each class previously defined. The vector is then assigned to the class that has the largest value of the discriminant function Q. This proceeds until all the radiance vectors for the imaged area are processed. One pitfall of this method is that a vector is always assigned to one of the classes even though it actually may not be similar to any of the classes. This problem may be attacked by a thresholding approach.

Since the $|\Sigma_i|^{1/2}$ and the Σ_i^{-1} need only be calculated once for each class, the most time consuming part of the calculation for each data vector is the quadratic computation of $(\overline{X} - \overline{\mu})^t \; \Sigma_i^{-1} \; (\overline{X} - \overline{\mu})$.

Thus the supervised method of classification uses statistics generated by a large number of samples to describe each class of data that a vector may be assigned to. Once these statistics are calculated, the discriminant function must be calculated for each class for every data vector. The vector is then assigned to one of those classes by inspection of the discriminant functions.

The ASTEP program has the supervised classification scheme described above implemented as a classification module. Training sets of data are usually located by comparing clustering outputs as described above with aerial photos or maps. The homogeniety of each training set may be tested by histograms of the data. Next, the statistics for each training class are computed and saved on magnetic tape. When the supervised module is requested, these training set statistics provide the necessary information to be able to classify other multispectral data into the selected classes.